



Predicting response to somatostatin analogues in acromegaly: machine learning-based high-dimensional quantitative texture analysis on T2-weighted MRI

Burak Kocak¹ · Emine Sebnem Durmaz² · Pinar Kadioglu³ · Ozge Polat Korkmaz³ · Nil Comunoglu⁴ · Necmettin Tanriover⁵ · Naci Kocer² · Civan Islak² · Osman Kizilkilic²

Received: 9 July 2018 / Revised: 10 October 2018 / Accepted: 7 November 2018
© European Society of Radiology 2018

Abstract

Objective To investigate the value of machine learning (ML)-based high-dimensional quantitative texture analysis (qTA) on T2-weighted magnetic resonance imaging (MRI) in predicting response to somatostatin analogues (SA) in acromegaly patients with growth hormone (GH)-secreting pituitary macroadenoma, and to compare the qTA with quantitative and qualitative T2-weighted relative signal intensity (rSI) and immunohistochemical evaluation.

Methods Forty-seven patients (24 responsive; 23 resistant patients to SA) were eligible for this retrospective study. Coronal T2-weighted images were used for qTA and rSI evaluation. The immunohistochemical evaluation was based on the granulation pattern of the adenomas. Dimension reduction was carried out by reproducibility analysis and wrapper-based algorithm. ML classifiers were k -nearest neighbours (k -NN) and C4.5 algorithm. The reference standard was the biochemical response status. Predictive performance of qTA was compared with those of the quantitative and qualitative rSI and immunohistochemical evaluation.

Results Five hundred thirty-five out of 828 texture features had excellent reproducibility. For the qTA, k -NN correctly classified 85.1% of the macroadenomas regarding response to SAs with an area under the receiver operating characteristic curve (AUC-ROC) of 0.847. The accuracy and AUC-ROC ranges of the other methods were 57.4–70.2% and 0.575–0.704, respectively. Differences in predictive performance between qTA-based classification and the other methods were significant ($p < 0.05$).

Conclusions The ML-based qTA of T2-weighted MRI is a potential non-invasive tool in predicting response to SAs in patients with acromegaly and GH-secreting pituitary macroadenoma. The method performed better than the qualitative and quantitative rSI and immunohistochemical evaluation.

Key Points

- Machine learning-based texture analysis of T2-weighted MRI can correctly classify response to somatostatin analogues in more than four fifths of the patients.
- Machine learning-based texture analysis performs better than qualitative and quantitative evaluation of relative T2 signal intensity and immunohistochemical evaluation.
- About one third of the texture features may not be excellently reproducible, indicating that a reliability analysis is necessary before model development.

Keywords Acromegaly · Growth hormone-secreting pituitary adenoma · Machine learning · Magnetic resonance imaging · Somatostatin

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00330-018-5876-2>) contains supplementary material, which is available to authorized users.

✉ Burak Kocak
drburakkocak@gmail.com

¹ Department of Radiology, Istanbul Training and Research Hospital, Istanbul, Turkey

² Department of Radiology, Cerrahpasa Medical Faculty, Istanbul University-Cerrahpasa, Istanbul, Turkey

³ Department of Endocrinology and Metabolism, Cerrahpasa Medical Faculty, Istanbul University-Cerrahpasa, Istanbul, Turkey

⁴ Department of Pathology, Cerrahpasa Medical Faculty, Istanbul University-Cerrahpasa, Istanbul, Turkey

⁵ Department of Neurosurgery, Cerrahpasa Medical Faculty, Istanbul University-Cerrahpasa, Istanbul, Turkey

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
AUC-ROC	Area under the receiver operating characteristic curve
GH	Growth hormone
ICC	Intra-class correlation coefficient
IGF-1	Insulin-like growth factor-1
<i>k</i> -NN	<i>k</i> -nearest neighbours
LoG	Laplacian of Gaussian
ML	Machine learning
MRI	Magnetic resonance imaging
qTA	Quantitative texture analysis
ROI	Region of interest
rSI	Relative signal intensity
SA	Somatostatin analogue
SD	Standard deviation
WEKA	Waikato Environment for Knowledge Analysis

Introduction

Somatostatin analogues (SA) are widely used in the medical treatment of acromegaly patients with growth hormone (GH)-secreting pituitary adenoma [1, 2]. Nonetheless, they cannot ensure biochemical control of the disease, improvement of clinical symptoms, and tumour shrinkage in all patients [1]. Several predictors of SA response including clinical, biochemical, radiological, and histopathological features, and somatostatin receptors have been described [1]. Prediction of response to SAs is essential because of their high cost. Magnetic resonance imaging (MRI) is the imaging method of pituitary adenomas [3], and hypointensity on T2-weighted images may signify a better response to SA treatment [4–7].

Quantitative texture analysis (qTA) evaluates lesion patterns that may not be visually perceptible [8]. Machine learning (ML) covers of a broad range of advanced statistical algorithms used in building autonomous predictive models in response to training data. One study has reported the qTA of GH-secreting pituitary macroadenomas for predicting response to SA [9], however with a first-order histogram analysis and no validation.

The purpose of this study was to investigate the potential value of ML-based high-dimensional qTA on T2-weighted MRI in predicting the response of GH-secreting pituitary macroadenomas to SAs, and to compare the qTA with relative signal intensity (rSI) and immunohistochemical granulation pattern evaluation that may be related with response to SAs.

Materials and methods

Ethics

This retrospective study was approved by our institutional review board. The requirement for informed consent was waived.

Patients

We reviewed our institutional databases for acromegaly patients between January 2009 and December 2017. Our inclusion criteria were as follows: (i) patients with biochemical acromegaly diagnosis based on age-adjusted serum insulin-like growth factor-1 (IGF-1) level and GH level (GH nadir > 1 µg/L) following oral glucose tolerance test; (ii) patients with no biochemical remission (GH level > 1 µg/L or elevated age-adjusted IGF-1) 3 months after surgery; (iii) patients with histopathologically confirmed GH-secreting macroadenoma (≥ 10 mm); and (iv) patients with preoperative and pretreatment (with SA) pituitary MRI including coronal T2-weighted sequences performed in our institution. Our exclusion criteria were as follows: (i) semi-solid macroadenomas with a solid component having a maximum diameter less than 10 mm (some texture features require a sufficient volume or area) and (ii) patients with pituitary apoplexy (to avoid possible distortion in texture feature parameters).

MRI technique

MRI was performed using a 1.5-T unit (Siemens, MAGNETOM Avanto). We used only turbo spin-echo T2-weighted coronal images which is a standard [5, 6, 9]. The settings were as follows: TR, 2090 ms; TE, 104 ms; echo train length or turbo factor, 24; slice thickness, 2.5 mm; slice spacing, 2.8 mm; field of view, 180 × 180 mm; and matrix size, 224 × 320, allowing a pixel size from 0.5 to 0.8 mm.

Image processing

The most important steps of the ML-based qTA are summarised in Fig. 1.

T2-weighted images underwent N4 bias field correction to remove low-frequency intensity non-uniformity [10].

To minimise differences, all data sets were normalised by centring the voxel image intensity values at the mean with the standard deviation (SD), known as the ± 3 sigma technique [11]. Image normalisation was done for all grey-level values in the image, not just for the segmentation. Normalisation was based on the formula:

$$f(x) = \frac{x - \mu(x)}{\sigma(x)}$$

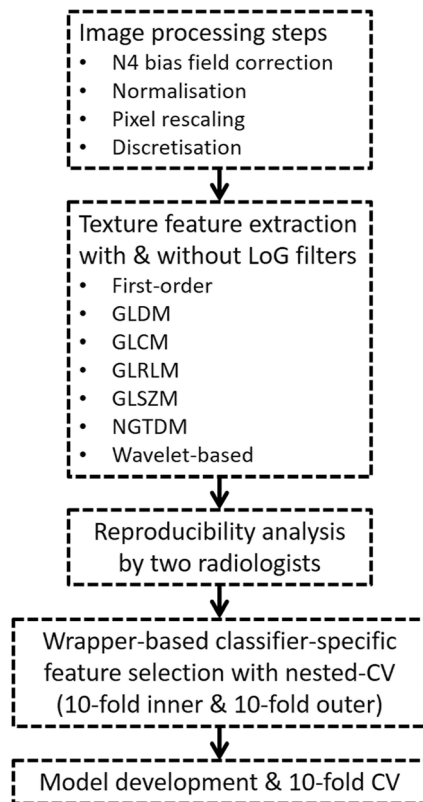


Fig. 1 Simplified flowchart showing the machine learning-based quantitative texture analysis pipeline. LoG, Laplacian of Gaussian; GLDM, grey-level dependence matrix; GLCM, grey-level co-occurrence matrix; GLRLM, grey-level run-length matrix; GLSZM, grey-level size zone matrix; NGTDM, neighbouring grey-tone difference matrix; CV, cross-validation

where $f(x)$ is normalised image intensity, x is original image intensity, $\mu(x)$ is mean image intensity value, and $\sigma(x)$ is the SD of the image intensity.

Pixel spaces in all image slices were rescaled to an in-plane resolution of $1 \times 1 \text{ mm}^2$ because the comparison of texture features necessitates identical spatial resolution [12]. Slice thickness was not rescaled because it was homogeneous.

The grey-level discretisation was done in the matrix representation of the grey levels in the segmentation, leaving the voxels outside segmentation unchanged. The discretisation was based on the following mathematical formula:

$$X_{b,i} = \left\lceil \frac{X_{gl,i}}{W} \right\rceil - \left\lfloor \frac{\min(X_{gl,i})}{W} \right\rfloor + 1$$

where $X_{b,i}$ is grey-level intensity after discretisation, $X_{gl,i}$ is grey-level intensity before discretisation, and W is the bin-width value, which was 0.06 in this study, corresponding to maximum 100 discrete grey-levels.

The N4 bias field correction, pixel resampling, normalisation, and discretisation were done with 3D data.

Except for the N4 bias field correction, all other processing steps (normalisation, resampling, and discretisation) were

done before the texture feature extraction using the same module of the software, not affecting the segmentation process with changes such as blurring of the image.

Texture feature extraction

Texture features were extracted using ‘SlicerRadiomics’ extension (Revision 8e5f1e8) of 3D-Slicer software (version 4.8.1) based on the Python package named ‘PyRadiomics’ [13]. The macroadenomas were independently segmented slice-by-slice (3D whole tumour segmentation) by two radiologists. To avoid partial volume effect, segmentation was performed excluding the peripheral tumour tissue, 1 mm from the visible lesion contour as well as the most anterior and posterior slices that included the lesion (Fig. 2). Although the lesions were segmented slice-by-slice, we forced the software package to perform the analysis only in the coronal plane because of the anisotropy of the coronal T2-weighted images (voxel size = $1 \times 1 \times 2.5 \text{ mm}^3$; slice spacing = 2.8 mm). The reason behind the slice-by-slice (3D) segmentation was to provide enough two-dimensional (2D) texture data by increasing the 2D segmentation area. 2D texture features using the entire tumour volume were extracted from both the original, filtered, and wavelet transformed images. Laplacian of Gaussian (LoG) filter was used for image filtration with values of 2 mm, 4 mm, and 6 mm (representing fine, medium, and coarse patterns). Of note, the LoG filtering and wavelet transformation were done to 3D volumetric data. The total number of the features extracted was 828 per lesion. Detailed texture feature groups are presented in [Online Supplement Part E1](#).

Dimension reduction

Two radiologists, blinded to the response status, independently segmented tumours slice-by-slice (3D whole tumour segmentation). Intra-class correlation coefficient (ICC) values were calculated for each texture feature using SPSS version 20. The features with an ICC value of ≥ 0.8 indicating ‘excellent’ reproducibility were included in the further analysis.

The wrapper-based classifier-specific feature selection and model optimisation were performed using WEKA toolkit version 3.8.2 (University of Waikato) [14, 15]. A nested cross-validation method with 10-fold inner and 10-fold outer loops was adopted (Fig. 3) [16, 17]. Details regarding the feature selection are presented in [Online Supplement Part E2](#).

Relative signal intensity evaluation

The rSI was evaluated qualitatively (visual) and quantitatively (with 2D region of interest (ROI) and 3D whole tumour segmentation).

The rSI of the adenoma was classified as follows: (i) hypointense (equal or less than the white matter of the

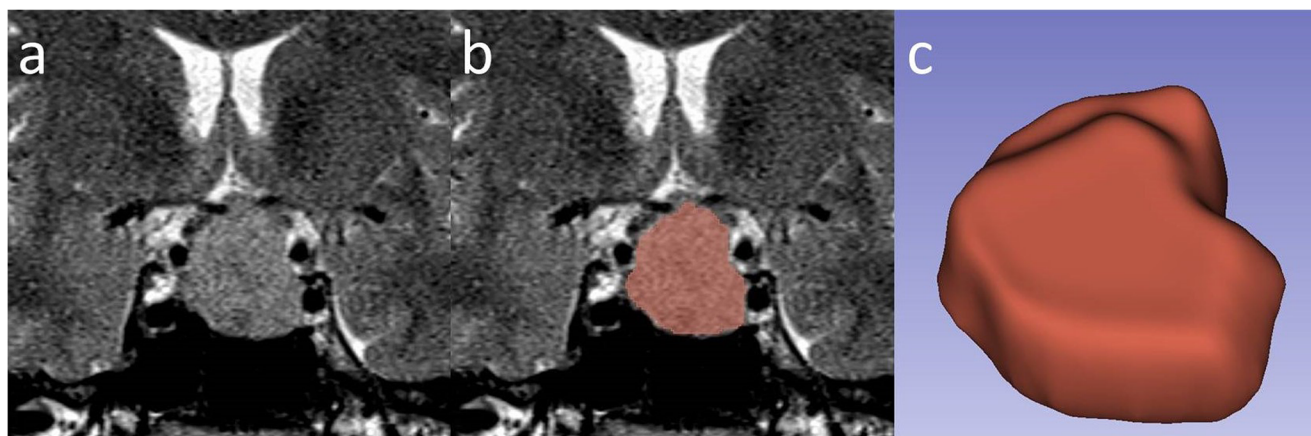


Fig. 2 Three-dimensional (3D) whole tumour segmentation for quantitative texture analysis and 3D segmentation-based quantitative relative signal intensity evaluation. (a) A hyperintense macroadenoma

with small patchy hypointense foci in the coronal T2-weighted image. (b, c) Slice-by-slice segmentation and 3D modelling of the macroadenoma

temporal lobe); (ii) hyperintense (equal or higher than the grey matter); and (iii) isointense (between white and grey matter) [5]. They were further grouped as follows: (i) T2-hypointense versus (ii) others for statistics.

For ROI-based quantitative rSI, the mean signal intensity was measured on two consecutive coronal T2-weighted images from the largest solid portion of the adenoma, white and grey matters of temporal lobe (Fig. 4) [5]. We also used the same 3D segmentation data used in qTA for rSI evaluation to allow comparison.

The qualitative (visual) rSI evaluation was done by two radiologists. In case of disagreement, the final decision was reached by consensus.

Immunohistochemical evaluation

Based on the staining characteristics using monoclonal cytokeratin antibody, the macroadenomas were divided

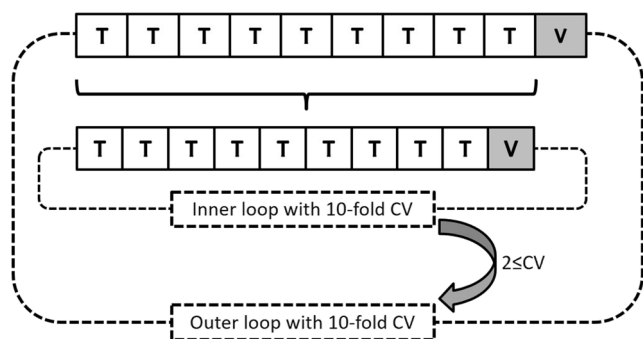


Fig. 3 Nested cross-validation with 10-fold inner loop and 10-fold outer loop. For each outer fold, the inner loop runs 10-fold cross-validation. The texture features having at least two cross-validations in the inner loop move to the outer fold. The 10-fold in the outer loop corresponds to the regular 10-fold cross-validation used in model development and validation. On the other hand, the 10-fold in the inner loop corresponds to the actual feature selecting process. Hence, this process creates ten different combinations of training and validation split. T, training; V, validation; CV, cross-validation

into three groups as follows: (i) densely; (ii) transitionally; and (iii) sparsely granulated [18]. Because sparsely granulated adenomas are considered having a poor SA response [1, 18], the final groups were (i) sparsely granulated and (ii) the others.

Response and resistance criteria

The reference standard was biochemical response to SA treatment. Three months after surgery, SA treatment was initiated for patients with a GH level $> 1 \mu\text{g/L}$ or elevated age-adjusted IGF-1. Patients were considered resistant if GH or age-adjusted IGF-1 levels were still elevated after 6 months of therapy with octreotide (40 mg per 28 days) or lanreotide (120 mg per 28 days).

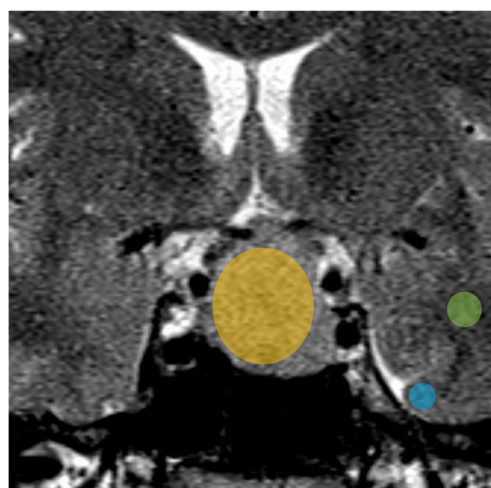


Fig. 4 Region of interests (ROI) used in ROI-based quantitative relative signal intensity evaluation. The ROIs are placed on the largest solid portion of the adenoma (yellow), temporal white matter (green), and temporal grey matter (blue). Please note that ROIs are drawn for two consecutive slices

Statistical analysis

The ML-based classifications were performed using WEKA toolkit version 3.8.2. The *k*-nearest neighbours (*k*-NN) classifier (IBk in WEKA toolkit) was utilised in qTA-based classifications [19]. The search algorithm for *k*-NN was linear with a Euclidean distance function. To minimise potential overfitting, we created models with five-nearest neighbours (5-NN). The C4.5 decision tree classifier was utilised in rSI-based classifications. In the WEKA toolkit, the C4.5 algorithm is represented with J48 [20]. The C4.5 (or the J48 in WEKA) is a simple ML scheme. We used this classifier for a binary classification problem (presence or absence of the response to SAs). The primary purpose of using this algorithm to obtain similar performance metrics from the software to create comparable metrics for the *k*-NN.

For qTA, a 10-fold cross-validation procedure was adopted for the validation of the model, calculating performance metrics by averaging these ten different validation performances. On the other hand, the models for the other methods were created on the whole data, resulting in a single performance metric.

The main performance evaluation metric was the area under the receiver operating characteristic curve (AUC-ROC) [21]. In addition, sensitivity, specificity, precision (positive predictive value), recall, *F*-measure, the Matthews correlation coefficient, and the area under the precision-recall curve were calculated as well. Comparisons of the AUC-ROCs derived from the qTA (10-fold cross-validated), qualitative and quantitative rSI, and immunohistochemical evaluations (single AUC-ROC value for each method) were performed using the one-sample Wilcoxon signed-ranks test [22].

Shapiro-Wilk test was used for the assessment of normal distribution. The difference of the mean signal intensity between 2D ROI and 3D segmentation data was analysed using the paired *t* test.

Cohen's kappa (*k*) was run to determine the strength of agreement between two observers' judgments on qualitative rSI evaluation. Interobserver agreement was judged as according to the following rating: 0.00–0.20 = slight; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial; and 0.81–1.00 = excellent.

Results

Patient demographics and characteristics

Forty-seven patients with acromegaly and histopathologically proven GH-secreting macroadenoma were included in the analysis. The patient demographics and characteristics are presented in Table 1.

3D segmentation and ROI characteristics

Mean (SD) of the 3D segmentation volume was 4734.1 mm³ (10880), of the maximum segmentation diameter was 24.3 mm (12.2), and of the ROI area was 95.23 mm² (114.02).

Mean (SD) signal intensity of the ROI-based was 271.85 (78.05) and of 3D segmentation-based was 191.02 (57.49). They were statistically different (*p* < 0.05).

Table 1 Patient characteristics and demographics

Patient characteristics	Value
Age	
Mean ± SD	37.7 ± 9.9 years
Range	19–59 years
Gender	
Female	28 patients
Male	19 patients
GH level	
Mean ± SD	23.7 ± 23.2 µg/L
Median	14.3 µg/L
Interquartile range	29.2 µg/L
IGF-1 level	
Mean ± SD	767.9 ± 287.8 mmol/L
Median	731.5 mmol/L
Interquartile range	336.2 mmol/L
Tumour volume*	
Mean ± SD	4734.1 ± 10,880 mm ³
Median	2057.8 mm ³
Interquartile range	4038.3 mm ³
Maximum tumour diameter	
Mean ± SD	24.3 ± 12.2 mm
Median	21.8 mm
Interquartile range	13.1 mm
Preoperative SA treatment	
Yes	13 patients
No	34 patients
Response to SA	
Responsive	24 patients
Resistant	23 patients
Granulation pattern	
Sparse	27 patients
Others**	20 patients

*Also represent the segmentation volume used for quantitative texture analysis

**Dense or transitional

SD, standard deviation; GH, growth hormone; IGF-1, insulin-like growth factor-1; SA, somatostatin analogue

Reproducibility analysis

Following reproducibility analysis by two radiologists, 293 out of 828 features were excluded based on the predefined ICC cutoff value ($ICC < 0.8$). The remaining 535 were included in further analysis.

Wrapper-based feature selection

In the initial run, the wrapper-based classifier-specific feature selection algorithm yielded 12 texture features. In the following runs, the number of texture features decreased to four ([Online Supplement Part E3](#)). Distributions of the selected feature values between responsive and resistant groups are presented in Figs. 5 and 6.

qTA-based classification

Using the selected features, the k -NN algorithm correctly classified 85.1% (40 out of 47) of the patients regarding response status to SAs with an AUC-ROC value of 0.847. Each AUC-ROC value in the 10-fold cross-validation is presented in Table 2. For detecting the responsive group, the sensitivity, specificity, and precision (or positive predictive value) were 87.5%, 82.6%, and 84%, respectively. For detecting the resistant group, the sensitivity, specificity, and precision were 82.6%, 87.5%, and 86.4%.

Quantitative rSI-based classification

In the 2D ROI-based rSI evaluation of the macroadenomas (10 macroadenomas T2-hypointense; 37 T2-isointense or

hyperintense), the C4.5 correctly classified 57.4% (27 out of 47) of the macroadenomas regarding response status with an AUC-ROC of 0.581.

In the 3D segmentation-based rSI evaluation of the macroadenomas (22 macroadenomas T2-hypointense; 25 T2-isointense or hyperintense), the C4.5 correctly classified 57.4% (27 out of 47) of the macroadenomas regarding response status with an AUC-ROC of 0.575.

Qualitative (visual) rSI-based classification

Interobserver agreement between two observers was substantial ($kappa$ (k) coefficient = 0.651).

Using visual rSI method and consensus data (17 macroadenomas T2-hypointense; 30 T2-isointense or hyperintense), the C4.5 correctly classified 59.6% (28 out of 47) of the macroadenomas regarding response status with an AUC-ROC of 0.599.

Granulation pattern-based classification

Based on the immunohistochemical granulation pattern (27 macroadenomas sparsely granulated; 20 densely or transitionally granulated), the C4.5 correctly classified 70.2% (33 out of 47) of the macroadenomas regarding response status with an AUC-ROC of 0.704.

qTA versus other methods

Considering the AUC-ROC performance metric (10-fold cross-validation values for qTA; single value for the other methods), there were significant differences between (i) qTA

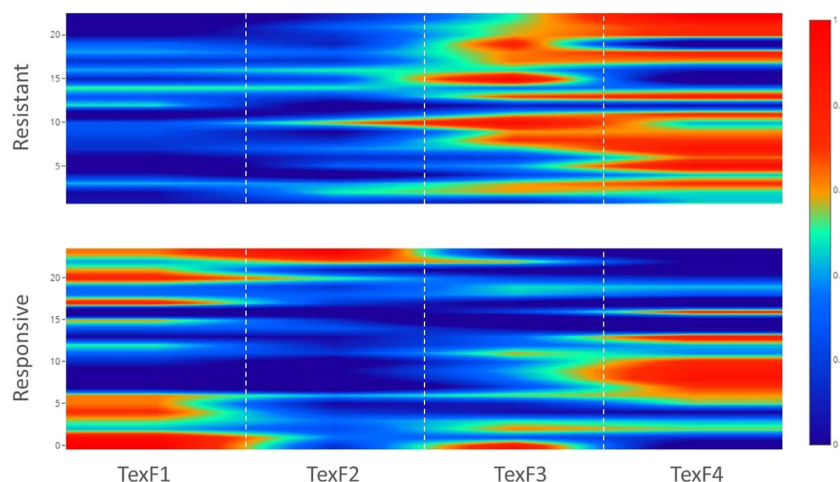


Fig. 5 Smoothed heat map created using the selected subset of texture features from all patients in the study. The map shows the distribution of normalised (0 to 1) texture feature values between responsive and resistant groups. Changes in colours and their shades indicate a difference in texture feature values in and between groups. TexF1, grey-level co-occurrence matrix (GLCM) Idn (inverse difference normalised)

in the image with a LoG filter of 2 mm; TexF2, the first-order maximum in the image with a LoG filter of 6 mm; TexF3, the first-order median in the image with wavelet energy in low/high-frequency bands; TexF4, neighbouring grey-tone difference matrix (NGTDM) coarseness in the original image

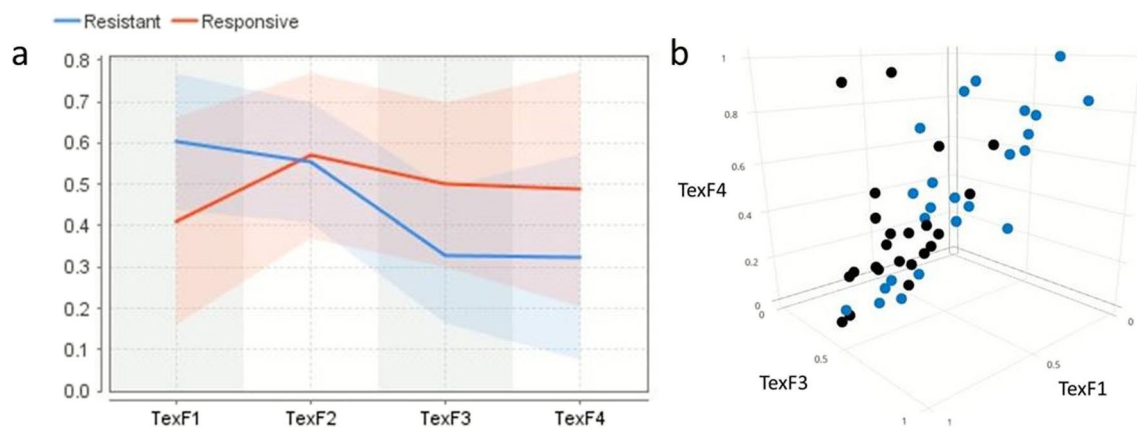


Fig. 6 (a) Deviation plot created with normalised values (0 to 1) of texture features showing the degree of overlap between responsive and resistant groups. Significant overlap is visually apparent in TexF2. Please note that although TexF2 has significant overlap, it still makes a positive contribution to the model's predictive accuracy. (b) A three-dimensional (3D) scatter plot created using least overlapping features with their normalised values (0 to 1) shows the individual place of the features in 3D

space. Blue circles, responsive group; black circles, resistant group; TexF1, grey-level co-occurrence matrix (GLCM) Idn (inverse difference normalised) in the image with a LoG filter of 2 mm; TexF2, the first-order maximum in the image with a LoG filter of 6 mm; TexF3, the first-order median in the image with wavelet energy in low/high-frequency bands; TexF4, neighbouring grey-tone difference matrix (NGTDM) coarseness in the original image

and 2D ROI-based quantitative rSI evaluation ($z = 2.8$; $p < 0.05$); (ii) qTA and 3D segmentation-based quantitative rSI evaluation ($z = 2.8$; $p < 0.05$); (iii) qTA and qualitative (visual) rSI evaluation ($z = 2.8$; $p < 0.05$); and (iv) qTA and granulation pattern-based evaluation ($z = 2.8$; $p < 0.05$).

Table 3 presents the performance metrics of all the methods.

Discussion

The most important finding was that the k -NN classifier correctly classified more than four fifths of the macroadenomas. The predictive performance of the ML-based qTA was better than those of quantitative and qualitative rSI, and immunohistochemical granulation pattern evaluation.

Table 2 Each area under the receiver operating characteristic curve (AUC-ROC) value in the 10-fold cross-validation of the machine learning-based quantitative texture analysis

Folds	AUC-ROC
1	0.843
2	0.842
3	0.848
4	0.848
5	0.815
6	0.855
7	0.875
8	0.853
9	0.855
10	0.832

AUC-ROC, area under the receiver operating characteristic curve

The literature suggests that preoperative SA treatment improves the surgical outcomes [23–26]. Resistance to SA treatment may delay surgery and deteriorate the surgical outcomes. Hence, predicting response or resistance with preoperative predictors or biomarkers is important. Non-invasive ML-based qTA might be an interesting method.

There has been only one study of qTA against rSI in predicting response to SA treatment [9]. The authors used first-order histogram analysis with very few texture features and reported that the overall diagnostic accuracy of the histogram-based model was 82.4% for predicting good response with an AUC-ROC value of 0.861. However, there was no validation. Furthermore, they reported that the predictive performance of the histogram-based method was not different from that of the visual T2-weighted intensity evaluation. In our analysis, none of the first-order features obtained from the original image was selected by the feature selection algorithm. Conversely, some first-order features extracted from filtered or transformed images were selected. Using a higher number of features and internal validation, we found that prediction by qTA was superior to rSI evaluation.

Regarding the definition and interpretation of the selected features for model development, TexF1 corresponds to local homogeneity in finely filtered images. TexF2 corresponds to the maximum signal intensity in coarsely filtered images. TexF3 indicates median signal intensity in low- and high-frequency decomposition images. TexF4 corresponds to spatial intensity changes. According to TexF1 and TexF4, the responsive macroadenomas to SAs were locally more homogeneous in finely filtered images and more non-uniform in the original images.

Considering our small patient population, our classifier was evaluated with a complex nested cross-validation approach

Table 3 Performance of quantitative texture analysis, quantitative relative signal intensity evaluation (ROI-based and 3D segmentation-based), qualitative (visual) relative signal intensity evaluation, and immunohistochemical granulation pattern-based evaluation in predicting response to somatostatin analogues

Methods	Sensitivity	Specificity	Precision	Recall	<i>F</i> -measure	MCC	AUC-ROC	AUC-PRC	Confusion matrix		
									N	Y	
Quantitative TA											
Resistant	82.6%	87.5%	86.4%	82.6%	0.844	0.702	0.847	0.777	19	4	No
Responsive	87.5%	82.6%	84%	87.5%	0.857			0.828	3	21	Yes
ROI-based quantitative rSI											
Resistant	87%	70.8%	54.1%	87%	0.667	0.197	0.581	0.534	20	3	No
Responsive	70.8%	87%	70%	29.2%	0.412			0.566	17	7	Yes
3D quantitative rSI											
Resistant	60.9%	54.2%	56%	60.9%	0.583	0.151	0.575	0.532	14	9	No
Responsive	54.2%	60.9%	59.1%	54.2%	0.565			0.554	11	13	Yes
Qualitative (visual) rSI											
Resistant	73.9%	45.8%	56.7%	73.9%	0.642	0.205	0.599	0.547	17	6	No
Responsive	45.8%	73.9%	64.7%	45.8%	0.537			0.573	13	11	Yes
Granulation pattern											
Resistant	78.3%	62.5%	66.7%	78.3%	0.720	0.412	0.704	0.628	18	5	No
Responsive	62.5%	78.3%	75%	62.5%	0.682			0.660	9	15	Yes

N and Y indicate classification results. No and Yes indicate reference standard

TA, texture analysis; rSI, relative signal intensity evaluation; ROI, region of interest; 3D, three-dimensional; MCC, Matthews correlation coefficient; AUC-ROC, area under the receiver operating characteristic curve; AUC-PRC, area under the precision-recall curve; N and No, resistant to somatostatin analogues; Y and Yes, responsive to somatostatin analogues

[27, 28]. It reduces the bias and gives a similar estimate of the error to that of independent validation [28]. Whole data might also be considered but might have led to bias due to the use of the same data set for feature selection and model development, also called ‘double-dipping phenomenon’ [16]. A random split of the data, creating separate training and validation data set, could mimic external validation but in such small data sets, the chance factor can deeply affect the results.

Generalisation of these results is subject to several limitations. The number of patients was small considering the numerous texture features. We needed to exclude patients with post-surgical biochemical remission as it could be related to surgery. The risk of over-fitting is an important issue, however the cross-validation technique intended to minimise it [16, 17]. In addition, we used five-nearest neighbours (5-NN) for the same purpose. We could have used each 2D segmentation in order to increase the number of labelled data. However, considering the very small size of tumours, this might have hampered texture analysis. In spite of a uniform imaging protocol, slight differences are unavoidable in a retrospective study. We applied N4 bias field correction [10], normalisation [11], discretisation [12], and pixel rescaling [12] to minimise differences. Although one fourth of our patients had preoperative SA treatment, which could be seen as a bias, we only used preoperative and pretreatment MRI studies. The methods

shown here can only be applied to GH-secreting macroadenomas and cannot be extrapolated to others.

Conclusions

The results suggest that ML-based qTA on T2-weighted MRI has the potential to predict response to SAs in acromegaly patients with a GH-secreting pituitary macroadenoma, and performs better than quantitative and qualitative T2-weighted rSI, or immunohistochemical granulation pattern evaluation.

Funding The authors state that this work has not received any funding.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Burak Kocak, MD.

Conflict of interest The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors (Burak Kocak, MD) has significant statistical expertise.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Methodology

- retrospective
- diagnostic or prognostic study
- performed at one institution

References

- Paragliola RM, Corsello SM, Salvatori R (2017) Somatostatin receptor ligands in acromegaly: clinical response and factors predicting resistance. *Pituitary* 20:109–115. <https://doi.org/10.1007/s11102-016-0768-4>
- Melmed S, Bronstein MD, Chanson P et al (2018) A consensus statement on acromegaly therapeutic outcomes. *Nat Rev Endocrinol* 14:552–561. <https://doi.org/10.1038/s41574-018-0058-5>
- Bonneville JF, Bonneville F, Cattin F (2005) Magnetic resonance imaging of pituitary adenomas. *Eur Radiol* 15:543–548. <https://doi.org/10.1007/s00330-004-2531-x>
- Puig-Domingo M, Resmini E, Gomez-Anson B et al (2010) Magnetic resonance imaging as a predictor of response to somatostatin analogs in acromegaly after surgical failure. *J Clin Endocrinol Metab* 95:4973–4978. <https://doi.org/10.1210/jc.2010-0573>
- Heck A, Ringstad G, Fougner SL et al (2012) Intensity of pituitary adenoma on T2-weighted magnetic resonance imaging predicts the response to octreotide treatment in newly diagnosed acromegaly. *Clin Endocrinol (Oxf)* 77:72–78. <https://doi.org/10.1111/j.1365-2265.2011.04286.x>
- Shen M, Zhang Q, Liu W et al (2016) Predictive value of T2 relative signal intensity for response to somatostatin analogs in newly diagnosed acromegaly. *Neuroradiology* 58:1057–1065. <https://doi.org/10.1007/s00234-016-1728-4>
- Potorac I, Petrossians P, Daly AF et al (2016) T2-weighted MRI signal predicts hormone and tumor responses to somatostatin analogs in acromegaly. *Endocr Relat Cancer* 23:871–881. <https://doi.org/10.1530/ERC-16-0356>
- Lubner MG, Smith AD, Sandrasegaran K, Sahani DV, Pickhardt PJ (2017) CT texture analysis: definitions, applications, biologic correlates, and challenges. *Radiographics* 37:1483–1503. <https://doi.org/10.1148/rg.2017170056>
- Heck A, Emblem KE, Casar-Borota O, Bollerslev J, Ringstad G (2016) Quantitative analyses of T2-weighted MRI as a potential marker for response to somatostatin analogs in newly diagnosed acromegaly. *Endocrine* 52:333–343. <https://doi.org/10.1007/s12020-015-0766-8>
- Tustison NJ, Avants BB, Cook PA et al (2010) N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging* 29:1310–1320. <https://doi.org/10.1109/TMI.2010.2046908>
- Collewet G, Strzelecki M, Mariette F (2004) Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. *Magn Reson Imaging* 22:81–91. <https://doi.org/10.1016/j.mri.2003.09.001>
- Shafiq-Ul-Hassan M, Zhang GG, Latifi K et al (2017) Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys* 44:1050–1062. <https://doi.org/10.1002/mp.12123>
- van Griethuysen JJM, Fedorov A, Parmar C et al (2017) Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 77:e104–e107. <https://doi.org/10.1158/0008-5472.CAN-17-0339>
- Kohavi R, John GH (1997) Wrappers for feature subset selection. *Artif Intell* 97:273–324. [https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/10.1016/S0004-3702(97)00043-X)
- Bernejo P, Gamez JA, Puerta JM (2011) Improving incremental wrapper-based subset selection via replacement and early stopping. *Intern J Pattern Recognit Artif Intell* 25:605–625. <https://doi.org/10.1142/S0218001411008804>
- Mwangi B, Tian TS, Soares JC (2014) A review of feature reduction techniques in neuroimaging. *Neuroinformatics* 12:229–244. <https://doi.org/10.1007/s12021-013-9204-3>
- Cawley GC, Talbot NLC (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 11:2079–2107
- Obari A, Sano T, Ohyama K et al (2008) Clinicopathological features of growth hormone-producing pituitary adenomas: difference among various types defined by cytokeratin distribution pattern including a transitional form. *Endocr Pathol* 19:82–91. <https://doi.org/10.1007/s12022-008-9029-z>
- Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. *Mach Learn* 6:37–66. <https://doi.org/10.1007/BF00153759>
- Salzberg SL (1994) C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach Learn* 16:235–240. <https://doi.org/10.1007/BF00993309>
- Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit* 30:1145–1159. [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2)
- Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7:1–30
- Carlsen SM, Lund-Johansen M, Schreiner T et al (2008) Preoperative octreotide treatment in newly diagnosed acromegalic patients with macroadenomas increases cure short-term postoperative rates: a prospective, randomized trial. *J Clin Endocrinol Metab* 93:2984–2990. <https://doi.org/10.1210/jc.2008-0315>
- Mao ZG, Zhu YH, Tang HL et al (2010) Preoperative lanreotide treatment in acromegalic patients with macroadenomas increases short-term postoperative cure rates: a prospective, randomised trial. *Eur J Endocrinol* 162:661–666. <https://doi.org/10.1530/EJE-09-0908>
- Shen M, Shou X, Wang Y et al (2010) Effect of presurgical long-acting octreotide treatment in acromegaly patients with invasive pituitary macroadenomas: a prospective randomized study. *Endocr J* 57:1035–1044. <https://doi.org/10.1507/endocrj.K10E-203>
- Bacigaluppi S, Gatto F, Anania P et al (2016) Impact of pre-treatment with somatostatin analogs on surgical management of acromegalic patients referred to a single center. *Endocrine* 51:524–533. <https://doi.org/10.1007/s12020-015-0619-5>
- Kuhn M, Johnson K (2013) Over-fitting and model tuning. In: *Applied predictive modeling*. Springer New York, New York, pp 61–92
- Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7(91). <https://doi.org/10.1186/1471-2105-7-91>